Estimating the Bayes Error from Empirical Data

LDRD ER-CSSE

Name	Don Hush			**Z Number**	115295	**Group**	CIC-3
Phone	5-2722	**FAX**	5-5220	**Mail Stop**	B265	**E-mail**	dhush@lanl.gov

1 Overview

Pattern recognition is one of the most fundamental and important types of information processing. The *Bayes error* is the error rate achieved by the optimal classifier for a given pattern recognition problem. Therefore the Bayes error represents the best achievable performance for *any* classifier. The Bayes error is rarely known in practice because designing a classifier which achieves the Bayes error requires exact knowledge of the probability distributions from which the data are drawn. This makes it difficult for the practitioner to determine how close a classifier is to optimal, and consequently to decide how much effort should be devoted to classifier improvement. In addition, when classifier performance fails to achieve pre-specified requirements it is difficult to determine whether this is due to lack of optimality in the classifier or lack of information in the data. Perhaps surprisingly, the Bayes error can be estimated from a given data set *without* designing a classifier that achieves this error. The estimation technique consists of choosing a family of classifiers for which the difference between their classification error and the Bayes error can be analytically approximated. Then, by fitting the actual (data dependent) classification error for a series of classifiers from this family to this bias expression it is possible to compute an estimate of the Bayes error. This approach was originally proposed by Fukunaga [Fuk90], and has been used to produce Bayes error estimates for some simple real world problems. *We propose to develop robust and efficient algorithms for estimating the Bayes error from empirical data.*

2 Technical Background

Consider a two-class pattern recognition problem where the goal is to correctly assign patterns (*e.g.*, "measurements of a system") to one of the two classes. In most practical problems it is impossible to do this without making errors. This can be illustrated with the measurements height and weight and the classes male and female. While it is generally true that men are heavier and taller than women, they cannot be perfectly classified on this basis alone. For example, consider only the measurement height. Clearly there is an overlap in the distribution of heights for men and women. A similar situation exists with respect to weight. Combining height and weight may help reduce overlap but will not eliminate it. The extent of overlap limits the degree to which men and women can be separated on the basis of these measurements. The Bayes classification error is a measure of ambiguity due to the overlap between measurement distributions. Since this ambiguity is *irreducible* for a given problem, the Bayes error is the *smallest* classification error achievable by *any* classifier.

More formally, following [DH73], let the number of measurements made at any one time be $\mathtt{d}$, and let $\boldsymbol{x} = [x_1\ x_2\ \cdots\ x_{\mathtt{d}}]$, be a set of measurements belonging to some class. Each measurement vector $\boldsymbol{x}$ is called a *sample*, and a collection of samples is called a *sample set*. For simplicity consider a two-class problem where ω_0 and ω_1 are the class labels. Let $\mathcal{P}(\omega_i \mid \boldsymbol{x})$

denote the probability that $\boldsymbol{x}$ belongs to class ω_i. Assume we have a classifier that assigns a label for each $\boldsymbol{x}$. An error is made or the sample $\boldsymbol{x}$ is misclassified, if $\boldsymbol{x}$ is assigned to the class ω_i when it is actually a member of some other class ω_j. The average classification error ϵ is

$$\epsilon = \int_{\boldsymbol{x}\to\omega_1} \mathcal{P}(\omega_0 \mid \boldsymbol{x})\, d\boldsymbol{x} + \int_{\boldsymbol{x}\to\omega_0} \mathcal{P}(\omega_1 \mid \boldsymbol{x})\, d\boldsymbol{x}. \tag{1}$$

In this equation the first term is the probability of assigning $\boldsymbol{x}$ the label ω_1 when its true label is ω_0 and the second term is the probability of assigning $\boldsymbol{x}$ the label ω_0 when its true label is ω_1. It is easy to show that the classification rule that minimizes ϵ is to assign $\boldsymbol{x}$ to the class ω_i with the highest probability $\mathcal{P}(\omega_i \mid \boldsymbol{x})$. The corresponding value of ϵ is called the Bayes error and denoted ϵ^*. These ideas are illustrated in Figure 1 for a 2-class problem with 1 measurement. The two regions $\boldsymbol{x} \to \omega_0$ and $\boldsymbol{x} \to \omega_1$ show the range of sample values $\boldsymbol{x}$ which are assigned to

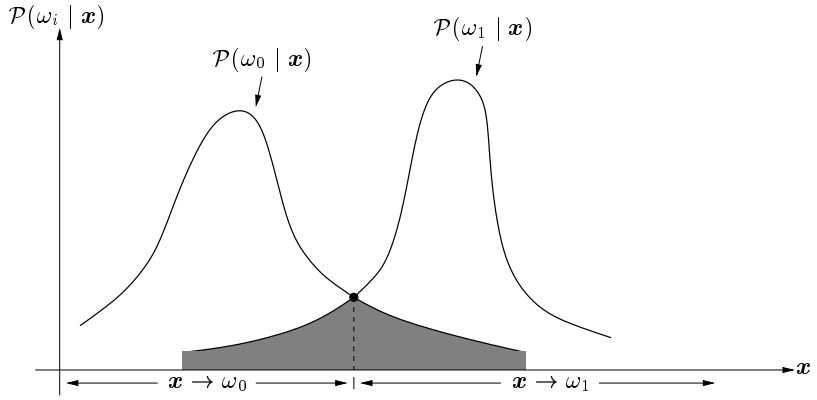

Figure 1: The decision regions of a Bayes classifier for a problem in which $\mathtt{m} = 2$ and $\mathtt{d} = 1$.

class ω_0 and ω_1 respectively. In the region $\boldsymbol{x} \to \omega_0$, the value of $\mathcal{P}(\omega_0 \mid \boldsymbol{x})$ is always greater than $\mathcal{P}(\omega_1 \mid \boldsymbol{x})$ and conversely in $\boldsymbol{x} \to \omega_1$. The area of the shaded region is the classification error ϵ^*. The dotted vertical line in the figure is at the point where $\mathcal{P}(\omega_0 \mid \boldsymbol{x}) = \mathcal{P}(\omega_1 \mid \boldsymbol{x})$ and represents the dividing line between the two regions.

In practice the the posterior probabilities $\mathcal{P}(\omega_i \mid \boldsymbol{x})$ are not known and so the goal is to determine a decision rule that best approximates the optimal one. This can be accomplished with a *discriminant function* $h(\boldsymbol{x})$ and a *classification threshold* τ that assigns $\boldsymbol{x}$ to class ω_1 when $h(\boldsymbol{x}) > \tau$ and to class ω_0 when $h(\boldsymbol{x}) < \tau$. Using Bayes rule to rewrite the conditional probabilities $\mathcal{P}(\omega_i \mid \boldsymbol{x})$, the discriminant function and classification threshold for a 2-class problem take the form

$$h(\boldsymbol{x}) = \frac{\hat{p}(\boldsymbol{x} \mid \omega_1)}{\hat{p}(\boldsymbol{x} \mid \omega_0)} \underset{\omega_0}{\overset{\omega_1}{\gtrless}} \frac{\hat{\mathcal{P}}(\omega_0)}{\hat{\mathcal{P}}(\omega_1)} = \tau, \tag{2}$$

where $\hat{p}(\cdot)$ and $\hat{\mathcal{P}}(\cdot)$ are estimates of the conditional densities and prior probabilities respectively.

For estimating the Bayes the error the family of k-nearest neighbor (kNN) classifiers was chosen because its classification error approaches the Bayes error asymptotically and its bias from the Bayes error can be analytically approximated. Given a finite set of labeled samples, $\mathcal{S} = \{(\boldsymbol{x}_i, \omega_i)\}$, a *volumetric k-nearest neighbor* classifier is obtained by substituting a 0th order discrete local approximation for the conditional densities $\hat{p}(\boldsymbol{x} \mid \omega_i)$. More specifically the density

estimate at $\boldsymbol{x}$ is the fraction of samples in a local region surrounding $\boldsymbol{x}$ divided by the volume of the local region, where the size of the local region is determined by the distance to the kth nearest neighbor from class ω_i, in other words

$$\hat{p}\big(\boldsymbol{x} \mid \omega_i\big) = \frac{k-1}{\mathfrak{n}_i \, v_k^{\omega_i}\big(\boldsymbol{x}\big)}. \tag{3}$$

In this equation $\mathfrak{n}_i$ is the number of samples from class ω_i and $v_k^{\omega_i}\big(\boldsymbol{x}\big)$ is the volume of the local region surrounding $\boldsymbol{x}$. The use of $k-1$ rather than k in the numerator ensures that the density estimate is asymptotically unbiased. The nature of this density estimate around a single point $\boldsymbol{x}(j)$ is shown in Figure 2. Intuitively, at every sample the *kNN* density estimate

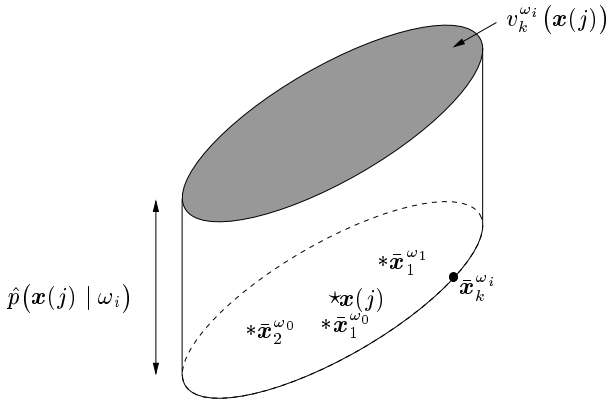

Figure 2: The *kNN* density estimate $\hat{p}\big(\boldsymbol{x}(j) \mid \omega_i\big)$ for the sample point $\boldsymbol{x}(j)$. The shaded region is the volume of the hyper-ellipsoid bounded by the kth closest point from class ω_i. The height of this window is the density estimate.

places a window of constant height which has a hyper-ellipsoidal footprint. The height of the window is computed from Equation (3). The size of the hyper-ellipsoid varies with the sample point being considered, and is determined by the distance from the sample point to the kth closest point from class ω_i, where this point is is denoted by $\bar{\boldsymbol{x}}_k^{\omega_i}$. The form of the distance measure determines the specific shape of the hyper-ellipsoid. Substituting Equation (3) in to Equation (2) for the 2-class problem and simplifying yields the *kNN* discriminant function

$$h_{kNN}(\boldsymbol{x}) = \frac{\mathcal{D}_0\big(\boldsymbol{x}, \bar{\boldsymbol{x}}_k^{\omega_0}\big)}{\mathcal{D}_1\big(\boldsymbol{x}, \bar{\boldsymbol{x}}_k^{\omega_1}\big)} \underset{\omega_0}{\overset{\omega_1}{\gtrless}} \left(\frac{\mathfrak{n}_1 \left|\hat{\boldsymbol{\Sigma}}_1\right|}{\mathfrak{n}_0 \left|\hat{\boldsymbol{\Sigma}}_0\right|} \frac{\hat{\mathcal{P}}(\omega_0)}{\hat{\mathcal{P}}(\omega_1)} \right)^{\frac{1}{\mathrm{d}}} = \tau_{kNN}. \tag{4}$$

In this equation $\mathcal{D}_i\big(\boldsymbol{x}, \bar{\boldsymbol{x}}_k^{\omega_i}\big) = (\boldsymbol{x} - \bar{\boldsymbol{x}}_k^{\omega_i})^T \hat{\boldsymbol{\Sigma}}_i^{-1} (\boldsymbol{x} - \bar{\boldsymbol{x}}_k^{\omega_i})$ is the *Mahalanobis distance* between the sample point $\boldsymbol{x}$ and its kth nearest neighbor from class ω_i, and $\hat{\boldsymbol{\Sigma}}_i$ is the estimate of the covariance matrix associated with the samples from class ω_i. In a *kNN* classifier the sample point $\boldsymbol{x}$ is assigned to class ω_0 if the scaled distance to the kth point from class ω_0 is less than the scaled distance to the kth point from class ω_1; otherwise it is assigned to class ω_1. The *kNN* discriminant function $h_{kNN}(\boldsymbol{x})$ is piecewise linear if $\hat{\boldsymbol{\Sigma}}_0 = \hat{\boldsymbol{\Sigma}}_1$ and piecewise quadratic otherwise, hence $h_{kNN}(\boldsymbol{x})$ is continuous. For a problem with two measurements per sample, the decision boundary $h_{1NN}(\boldsymbol{x}) = \tau_{1NN}$ for a *1NN* classifier is shown in Figure 3.

The statistical properties of *kNN* classifiers are well known and are discussed in [DGL96] and [Fuk90]. The *kNN* classifier is *consistent* in that its classification error approaches the

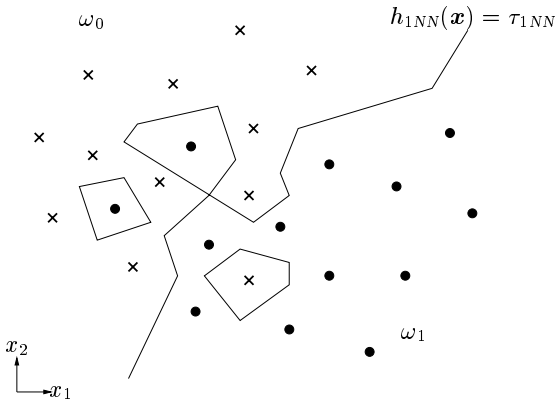

Figure 3: The decision boundary for a $1NN$ classifier for a problem with two measurements per sample where $\hat{\boldsymbol{\Sigma}}_0 = \hat{\boldsymbol{\Sigma}}_1$.

Bayes error asymptotically as both $k \to \infty$ and $\mathfrak{n} \to \infty$. However, for finite k and $\mathfrak{n}$, its classification error is biased away from the Bayes error. Fukunaga [Fuk90] has shown that this bias can be reduced by replacing τ_{kNN} with a threshold $\hat{\tau}$ that is chosen to minimize the classification error over the samples in $\mathcal{S}$. An approximation for the bias of the kNN classifier that is accurate to second order is

$$\epsilon_{kNN} \approx \epsilon^* + c_0(\Delta\tau)^2 + c_1\left(\frac{1}{k}\right) + (c_2 + c_3\Delta\tau)\left(\frac{k}{\mathfrak{n}}\right)^{\frac{2}{d}} + (c_4 + c_5\Delta\tau)\left(\frac{k}{\mathfrak{n}}\right)^{\frac{4}{d}}, \tag{5}$$

where $\Delta\tau = \hat{\tau} - \tau_{kNN}$ and $\mathfrak{n} = \mathfrak{n}_0 + \mathfrak{n}_1$. Note that $\epsilon_{kNN} = \epsilon^*$ as $k \to \infty$, $\mathfrak{n} \to \infty$, and $\hat{\tau} \to \tau_{kNN}$. Also note that Funkunaga's work shows, perhaps surprisingly, that for finite k and $\mathfrak{n}$ choosing $\hat{\tau} = \tau_{kNN}$ does *not* lead to the best estimates of ϵ^*. The bias approximation in Equation (5) is a function of only two independent variables; the neighborhood size k, and the threshold $\hat{\tau}$. In addition it is linear in seven undetermined coefficients; c_0, c_1, c_2, c_3, c_4, c_5 and the Bayes error ϵ^*. This suggests the following technique for estimating the Bayes error. Compute ϵ_{kNN} for $k = 1, 2, ...$k and then fit a curve to the resulting set of $\{(k, \epsilon_{kNN}(k))\}$ pairs to obtain the coefficients of Equation (5). This curve fit is illustrated in Figure 4. The zero order coefficient

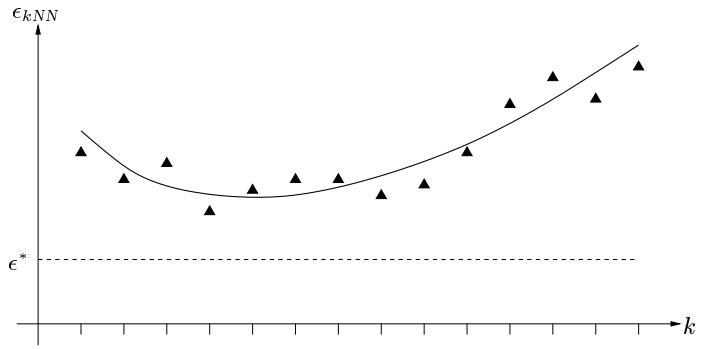

Figure 4: Curve fitting the bias of the volumetric kNN classifier.

from this fit is an estimate of the Bayes error! *Notice that this remarkably simple technique can be used to obtain accurate estimates of the Bayes error without building a classifier which achieves this error.*

We summarize by placing this idea in a historical context. Most of the work on Bayes error estimation has focused on estimating *bounds* for the Bayes error from data. It is shown in [DGL96] that even theoretically these bounds are not tight. Furthermore, it is shown in [FH87b] that in practice the estimates of these bounds often have large biases. Fukunaga has worked extensively on methods for tightening these bounds by reducing this bias [FH73, FF84, FF85, FH87b, FH87a, FH89, Fuk90]. In his later work Fukunaga [FH87a, Fuk90] introduced the idea of estimating the Bayes error directly by curve fitting a bias approximation (*i.e.*, Equation (5)), but his techniques are limited to two-class problems where the class distributions $p(\boldsymbol{x} \mid \omega_i)$ are unimodal. In addition, Fukunaga's work overlooks some important theoretical and practical issues that must be overcome to make this technique practically viable. *We propose to address these open issues and extend this technique to multi-class problems with multimodal distributions.*

3 Proposed Work and Significance to LANL

We propose to develop robust algorithms for estimating the Bayes error from large data sets. This requires fleshing out Fukunaga's proposed methodology in the following ways.

1. **Parameter Estimation**: Fukunaga develops several criteria for choosing $\hat{\tau}$ which we have explored in our previous work [HH99]. We will theoretically assess the advantages and disadvantages of these criteria, as well as investigating other possible criteria. We will then explore computationally efficient implementations of the various criteria. This will require the development of new algorithms that significantly extend Fukunaga's work. This issue must be addressed because good estimates of $\hat{\tau}$ are essential to obtaining good estimates of the Bayes error.

2. **Nearest Neighbor Computation**: The computational complexity of this approach is dominated by the the nearest neighbor computation. A brute force implementation is computationally expensive and will prohibit this method from scaling to large data sets. We will investigate data structures (*e.g.*, k-d trees) and algorithms for efficiently computing the nearest neighbors. In addition, we will explore recent methods, such as those in [IM98], for approximating nearest neighbor computations that promise to scale to extremely large problem sizes. Resolving these computational issues is critical in order to scale the Bayes error estimation procedure to data sets with a large number of samples and/or a large number of measurements per sample.

We also propose to extend Fukunaga's method in the following ways.

1. **Multiple Modes per Class**: Fukunaga's technique assumes that the conditional probability densities $p(\boldsymbol{x} \mid \omega_i)$ are unimodal. We will to extend this formalism to the case where $p(\boldsymbol{x} \mid \omega_i)$ are multimodal. This will require the use of a mixture model for $p(\boldsymbol{x} \mid \omega_i)$ and the incorporation of a maximum likelihood estimator for the parameters of this model, such as the EM algorithm and its recent extensions discussed by [McL92].

2. **Multiple Classes**: Fukunaga's procedure is designed for two-class problems. We will extend the method to cases where there are more than two classes. This will require the derivation of a new bias expression, since Equation (5) is valid only for the 2-class problem. Beyond this we believe that many of the lessons learned from implementing multiple modes per class can be leveraged to address problems with multiple classes.

Bayes error estimation can be applied to classification problems at LANL in areas such as computer security and weapons non-proliferation. Examples of computer security problems include document classification and network intrusion detection. Examples of problems in weapons non-proliferation include the detection of weapons tests and weapons factories. All of these problems are classification problems in the sense that the goal is to distinguish between a class of things that are defined to be "normal" and a class of things that are "abnormal". An important question in all of these problems is how well the classes can be distinguished using the currently measured quantities. An estimate of the Bayes error answers precisely this question. Another important issue in these problems is the determination of the optimal feature set. Features can be thought of as transformations of the data that filter out noise and retain the information essential for discrimination. In order to select features one must have some criteria that allows different features to be compared. The Bayes error estimate is one good candidate for such a criteria. To compare two features, transform the original sample data using each feature. Estimate the Bayes error for each of these two transformed sample sets. The better feature under this criteria is the one with the lower Bayes error estimate.

4 Specialist Reviewers

Tom Burr, NIS-7
Clint Scovel, CIC-3
Keinosuke Fukunaga, Purdue University, fukunaga@ecn.purdue.edu
Don Hummels, University of Maine, hummels@eece.maine.edu

5 Funding Breakout and Key Participants

Key Technical Staff: Don Hush (CIC-3), James Howse (X-8).

Funding: Funding is requested for 3 years. All funding is for labor of Don Hush and James Howse at approximately 0.5 FTE each per year. This leads to a total funding level of $170K per year.

References

[DGL96] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*, volume 31 of *Applications of Mathematics*. Springer-Verlag, Inc., New York, NY, 1996.

[DH73] R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis*. John Wiley & Sons, Inc., New York, NY, 1973.

[FF84] K. Fukunaga and T.E. Flick. An optimal global nearest neighbor metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(3):314–318, 1984.

[FF85] K. Fukunaga and T.E. Flick. The 2-NN rule for more accurate NN risk estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 7(1):107–112, 1985.

[FH73] K. Fukunaga and L.D. Hostetler. k-nearest-neighbor bayes-risk estimation. *IEEE Transactions on Information Theory*, 21(3):285–293, 1973.

[FH87a] K. Fukunaga and D.M. Hummels. Bayes error estimation using parzen and k-NN procedures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9(5):634–643, 1987.

[FH87b] K. Fukunaga and D.M. Hummels. Bias of nearest neighbor estimates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9(1):103–112, 1987.

[FH89] K. Fukunaga and D.M. Hummels. Leave-one-out procedures for nonparametric error estimates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(4):421–423, 1989.

[Fuk90] K. Fukunaga. *Introduction to Statistical Pattern Recognition.* Academic Press, Inc., San Diego, CA, 2nd edition, 1990.

[HH99] D.R. Hush and J.W. Howse. Predicting chilled hearth conditions: A classification approach. Technical Report LA–UR–99–6189, Los Alamos National Laboratory, 1999.

[IM98] P. Indyk and R. Motwanni. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the 30th Annual ACM Symposium on the Theory of Computing*, 1998.

[McL92] G.J. McLachlan. *Discriminant analysis and statistical pattern recognition.* John Wiley and Sons, Inc., New York, 1992.

Don R. Hush

Address

Group CIC-3	Phone: 505-665-2722
MS B265	FAX: 505-665-5220
Los Alamos National Laboratory	Email: dhush@lanl.gov
Los Alamos, NM 87545	

Education

PhD Elec Eng	1986	University of New Mexico
MS Elec Eng	1982	Kansas State University
BS Elec Eng	1980	Kansas State University (Summa Cum Laude)

Professional Experience

1998–present	Technical Staff Member, Los Alamos National Laboratory
1993–1998	Associate Professor, University of New Mexico
1987–1993	Assistant Professor, University of New Mexico
Summer 1991,1994,1996	Visiting Professor, Universidad de Vigo, Vigo, Spain
1986–1987	Technical Staff Member, Sandia National Laboratories
1994–1997	Associate Editor, IEEE Transactions on Neural Networks
1994–1998	Associate Editor, Signal Processing Magazine

Research Areas

Computational Learning Theory/Machine Learning/Pattern Recognition/Neural Networks
Numerical Optimization

Selected Recent Publications

M. Fugate, R. Christensen, D. Hush, and C. Scovel, "An equivalence relation between parallel calibration and principal component regression," submitted to *Journal of Chemometrics*, 2000.

D. Hush and C. Scovel, "Conditional performance bounds for machine learning," submitted to *Machine Learning*, 1999.

D. Hush and C. Scovel, "A new proof of concentration of Rademacher statistics," submitted to *Annals of Probability*, 1999.

D. Hush and C. Scovel, "On the VC dimension of bounded margin classifiers," to appear in Machine Learning, 2000.

D. Hush, "Training a sigmoid node is hard," *Neural Computation*, Vol. 11, pp. 1249–1260, 1999.

D. Hush and B. Horne, "Efficient algorithms for function approximation with piecewise linear sigmoidal networks," *IEEE Trans. Neural Networks*, Vol. 9, No. 6, pp. 1129–1141, 1998.

James W. Howse

Address

Group: X–8	Phone: 505-665-0603
Mail Stop: F645	FAX: 505-665-4479
Los Alamos National Laboratory	Email: jhowse@lanl.gov
Los Alamos, NM 87545	

Education

Ph.D. Electrical Engineering	University of New Mexico	1995
M.S. Electrical Engineering	University of Central Florida	1990
B.S. Engineering Physics	Lehigh University	1986

Professional Experience

1998–present	Technical Staff Member, Los Alamos National Laboratory
1996–1998	Postdoctoral Researcher, Los Alamos National Laboratory
1992	Research Assistant, NASA / Goddard Space Flight Center
1986–1987	Senior Staff Technologist, Bellcore

Research Areas

Support vector regression / Pattern classification / Numerical optimization

System identification and parameter estimation / Time series analysis and prediction

Tomographic image reconstruction and analysis / Numerical solutions of partial differential equations

Selected Recent Publications

"Solving a Thermal Regenerator Model using Implicit Newton-Krylov Methods", J.W. Howse, G.A. Hansen, D.J. Cagliostro, and K.R. Muske, Accepted for publication in *Numerical Heat Transfer*, January, 1999.

"Model-Based Control of a Thermal Regenerator. Part 1: Dynamic Model", K.R. Muske, J.W. Howse, G.A. Hansen, D.J. Cagliostro, Submitted to *Computers and Chemical Engineering*, August, 1999.

"Model-Based Control of a Thermal Regenerator. Part 2: Control and Estimation", K.R. Muske, J.W. Howse, G.A. Hansen, D.J. Cagliostro, Submitted to *Computers and Chemical Engineering*, August, 1999.

"Least Squares Estimation Techniques for Position Tracking of Radioactive Sources", J.W. Howse, L.O. Ticknor, and K.R. Muske, Submitted to *Automatica*, March, 1999.

"A Learning Algorithm for Applying Synthesized Stable Dynamics to System Identification", J.W. Howse, C.T. Abdallah, and G.L. Heileman, *Neural Networks*, Vol. 11, No. 1, pp. 81–87, 1998.